Week 16: Activity

Last Time

- Intro to Areal Data
- Areal Data Visualization
- Assessing Spatial Structure in Areal Data
- Spatial Smoothing with Areal Data

This Time

- Overview of Areal Data Models
- Simulating the fitting areal data

Areal Data Models: Disease Mapping

Areal data with counts is often associated with disease mapping, where there are two quantities for each areal unit: Y_i = observed number of cases of disease in county i and E_i = expected number of cases of disease in county i.

Note this can also be used to model generic count data on area units.

One way to think about the expected counts is

$$E_i = n_i \bar{r} = n_i \left(\sum_j y_j / \sum_j n_j \right),$$

where $\bar{r}$ is the overall disease rate and n_i is the population for region i.

However note that $\bar{r}$, and hence, E_i is a not fixed, but is a function of the data. This is called *internal standardization.*

An alternative is to use some standard rate for a given age group, such that $E_i = \sum_j n_{ij} r_j$. This is *external standardization.*

Often counts are assumed to follow the Poisson model where

$$Y_i | \eta_i \sim Poisson(E_i \eta_i),$$

where η_i is the relative risk of the disease in region i. This quantity is known as the *standardized morbidity ratio* (SMR).

Then the MLE of η_i is Y_i / E_i.

Consider a dataset with hospital admissions for respiratory disease in Glasgow. Plot both raw counts of hospital admissions and the SMR, which controlls for population.this areal data. Note that you can use leaflet or ggplot for this exercise.

```
data(respiratorydata)
data(GGHB.IZ)
```

Assess whether either the observed hospital counts or the standardized SMR show evidence of spatial structure.

Areal models with spatial structure

Poisson-lognormal models

The model can be written as

$$\begin{aligned} Y_i | \psi_i &\sim Poisson(E_i \psi_i) \\ log(\psi_i) &= x_i^T \beta + \theta_i + \phi_i \end{aligned}$$

where x_i are spatial covariates, θ_i corresponds to region wide heterogeneity (random effects), and ϕ_i captures local clustering (spatial structure).

Brook's Lemma and Markov Random Fields

To consider areal data from a model-based perspective, it is necessary to obtain the joint distribution of the responses

$$p(y_1, \ldots, y_n).$$

From the joint distribution, the *full conditional distribution*

$$p(y_i | y_j, j \neq i),$$

is uniquely determined.

Brook's Lemma states that the joint distribution can be obtained from the full conditional distributions.

When the areal data set is large, working with the full conditional distributions can be preferred to the full joint distribution.

More specifically, the response Y_i should only directly depend on the neighbors, hence,

$$p(y_i | y_j, j \neq i) = p(y_i | y_j, j \in \delta_i)$$

where δ_i denotes the neighborhood around i.

The idea of using the local specification for determining the global form of the distribution is Markov random field.

An essential element of a MRF is a *clique*, which is a group of units where each unit is a neighbor of all units in the clique

A *potential function* is a function that is exchangeable in the arguments. With continuous data a common potential is $(Y_i - Y_j)^2$ if $i \sim j$ (i is a neighbor of j).

A joint distribution $p(y_1, \ldots, y_n)$ is a Gibbs distribution if it is a function of Y_i only through the potential on cliques.

Conditional Autoregressive Models

Gaussian Model

Suppose the full conditionals are specifed as

$$Y_i | y_j, j \neq i \sim N\left(\sum_j b_{ij} y_j, \tau_i^2\right)$$

Then using Brooks' Lemma, the joint distribution is

$$p(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2} y^T D^{-1}(I - B) y\right),$$

where B is a matrix with entries b_{ij} and D is a diagonal matrix with diagonal elements $D_{ii} = \tau_i^2$.

The previous equation suggests a multivariate normal distribution, but $D^{-1}(I - B)$ should be symmetric.

Symmetry requires

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}, \ \forall \ i, j$$

In general, B is not symmetric, but setting $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$ satisfies the symmetry assumptions (given that we assume W is symmetric)

Now the full conditional distribution can be written as

$$Y_i | y_j, j \neq i \sim N\left(\sum_j w_{ij} y_j / w_{i+}, \tau^2 / w_{i+}\right)$$

Similarly the joint distribution is now

$$p(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\tau^2} y^T (D_w - W) y\right)$$

where D_w is a diagonal matrix with diagonal entries $(D_w)_{ii} = w_{i+}$

The joint distribution can also be re-written as

$$p(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\tau^2} \sum_{i \neq j} w_{ij} (y_i - y_j)^2\right)$$

However, both these formulations results in an improper distribution. This could be solved with a constraint, such as $Y_i = 0$.

The result is the joint distribution is improper, despite proper full conditional distributions. This model specification is often referred to as an *intrinsically autoregressive* model (IAR).

IAR

The IAR cannot be used to model data directly, rather this is used a prior specification and attached to random effects specified at the second stage of the hierarchical model.

The impropriety can be remedied by defining a parameter ρ such that $(D_w - W)$ becomes $(D_w - \rho W)$ such that this matrix is nonsingular.

The parameter ρ can be considered an extra parameter in the CAR model.

With or without ρ, $p(y)$ (or the Bayesian posterior when the CAR specification is placed on the spatial random effects) is proper.

When using ρ, the full conditional becomes

$$Y_i | y_j, j \neq i \sim N\left(\rho \sum_j w_{ij} y_j / w_{i+}, \tau^2 / w_{i+}\right)$$

Returning to the previously specified model

$$\begin{aligned} Y_i | \psi_i &\sim Poisson(E_i \psi_i) \\ log(\psi_i) &= x_i^T \beta + \theta_i + \phi_i \end{aligned}$$

when we place a CAR prior on ϕ and standard random effects on θ, this model is known as the Besag-York-Mollie (BYM) model.

specifically,

$$\begin{aligned} \phi_k | \phi_{-k}, W, \tau &\sim N(\frac{\sum_{i=1}^k w_{ki} \phi_i}{\sum_{i=1}^k w_{ki}}, \frac{\tau^2}{\sum_{i=1}^k w_{ki}}) \\ \theta_k &\sim N(0, \sigma^2) \end{aligned}$$

Alternatively we can specify the following model known as the Leroux model which uses the IAR framework where

$$\begin{aligned} Y_i | \psi_i &\sim Poisson(E_i \psi_i) \\ log(\psi_i) &= x_i^T \beta + \phi_i \\ \phi_k | \phi_{-k}, W, \tau &\sim N(\frac{\rho \sum_{i=1}^k w_{ki} \phi_i}{\rho \sum_{i=1}^k w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^k w_{ki} + 1 - \rho}) \end{aligned}$$

using `S.glm`, `S.CARbym`, and `S.CARleroux` from the `CARBayes` package, fit these spatial models and compare with information criteria.

Note that the above models result in a single smooth, spatial random surface (defined by the neighborhood structure). However, models can also be formulated to incorporate local spatial structure.

One option is the Lee and Mitchell approach, which models the w_{kj} terms rather than setting all to be zero or one. Fit this model using `S.CARdissimilarity` and compare to the previous models.